

University of Massachusetts Medical School

eScholarship@UMMS

Open Access Articles

Open Access Publications by UMMS Authors

2020-04-25

Transcriptomic profiling reveals extraordinary diversity of venom peptides in unexplored predatory gastropods of the genus *Clavus*

Aiping Lu
Tongji University

Et al.

Let us know how access to this document benefits you.

Follow this and additional works at: <https://escholarship.umassmed.edu/oapubs>



Part of the [Amino Acids, Peptides, and Proteins Commons](#), [Biochemical Phenomena, Metabolism, and Nutrition Commons](#), [Biochemistry Commons](#), [Bioinformatics Commons](#), [Computational Biology Commons](#), [Ecology and Evolutionary Biology Commons](#), [Genomics Commons](#), [Integrative Biology Commons](#), and the [Molecular Biology Commons](#)

Repository Citation

Lu A, Watkins M, Li Q, Robinson SD, Concepcion GP, Yandell M, Weng Z, Olivera BM, Safavi-Hemami H, Fedosov AE. (2020). Transcriptomic profiling reveals extraordinary diversity of venom peptides in unexplored predatory gastropods of the genus *Clavus*. Open Access Articles. <https://doi.org/10.1093/gbe/evaa083>. Retrieved from <https://escholarship.umassmed.edu/oapubs/4212>

Creative Commons License



This work is licensed under a [Creative Commons Attribution 4.0 License](#).

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Open Access Articles by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.

Transcriptomic profiling reveals extraordinary diversity of venom peptides in unexplored predatory gastropods of the genus *Clavus*

Aiping Lu¹, Maren Watkins², Qing Li^{3,4}, Samuel D. Robinson², Gisela P. Concepcion⁵, Mark Yandell^{3,6}, Zhiping Weng^{1,7}, Baldomero M. Olivera², Helena Safavi-Hemami^{8,9} and Alexander E. Fedosov^{10*}

¹ Department of Central Laboratory, Shanghai Tenth People's Hospital of Tongji University, School of Life Sciences and Technology, Tongji University, Shanghai, 200092, China

² Department of Biology, University of Utah, Salt Lake City, UT 84112, USA

³ Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA

⁴ High-Throughput Genomics and Bioinformatic Analysis Shared Resource, Huntsman Cancer Institute, University of Utah

⁵ Marine Science Institute, University of the Philippines-Diliman, Quezon City 1101, Philippines

⁶ Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT 84112, USA

⁷ Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA

⁸ Department of Biochemistry, University of Utah, Salt Lake City, UT 84112, USA

⁹ Department of Biology, University of Copenhagen, Copenhagen, Denmark

¹⁰ A.N. Severtsov Institute of Ecology and Evolution, Russian Academy of Science, 119071, Leninsky Prospect, 33, Moscow, Russia

* Author for Correspondence: Alexander E Fedosov, A.N. Severtsov Institute of Ecology and Evolution, Russian Academy of Science, 119071, Leninsky Prospect, 33, Moscow, Russia

Abstract

Predatory gastropods of the superfamily Conoidea number over 12,000 living species. The evolutionary success of this lineage can be explained by the ability of conoideans to produce complex venoms for hunting, defense and competitive interactions. Whereas venoms of cone snails (family Conidae) have become increasingly well studied, the venoms of most other conoidean lineages remain largely uncharacterized. In the present study we present the venom gland transcriptomes of two species of the genus *Clavus* that belong to the family Drilliidae. Venom gland transcriptomes of two specimens of *Clavus canalicularis*, and two specimens of *Cv. davidgilmouri* were analyzed, leading to the identification of a total of 1,176 putative venom peptide toxins (“drillipeptides”). Based on the combined evidence of secretion signal sequence identity, entire precursor similarity search (BLAST), and the orthology inference, putative *Clavus* toxins were assigned to 158 different gene families. The majority of identified transcripts comprise signal, pro-, mature peptide, and post- regions, with a typically short (<50 amino acids) and cysteine-rich mature peptide region. Thus drillipeptides are structurally similar to conotoxins. However, convincing homology with known groups of *Conus* toxins was only detected for very few toxin families. Among these are *Clavus* counterparts of *Conus* venom insulins (drillinsulins), porins (drilliporins), highly diversified lectins (drillilectins). The short size of most drillipeptides and structural similarity to conotoxins was unexpected, given that most related conoidean gastropod families (Terebridae and Turridae) possess longer mature peptide regions. Our findings indicate that, similar to conotoxins, drillipeptides may represent a valuable resource for future pharmacological exploration.

Keywords: Conoidea, venom, *Clavus*, drillipeptides, venom gland, transcriptome

Introduction

The predatory gastropods of the superfamily Conoidea J. Fleming, 1822 are a megadiverse taxon of marine mollusks, whose rapid evolutionary radiation was made possible by the development of a complex venom apparatus in ancestors of recent conoideans (Bouchet et al. 2011; Abdelkrim et al. 2018). Conoidean venoms are complex and contain potent and often highly selective peptides that are used for prey capture and defense. It is widely accepted that the diversification of Conoidea was underpinned by evolution of both their venoms and venom delivery systems (Shimek & Kohn, 1981; Olivera et al. 1985; Taylor et al. 1993; Dutertre et al. 2015). The elegant, uniquely efficient envenomation mechanism of many conoideans, wherein the venom is injected into prey through a hollow hypodermic harpoon – a highly specialized type of the molluscan grazing organ, the radula – is best known from fish-hunting cone snails in the genus *Conus* Linnaeus, 1758 (Kantor & Taylor 2000; Olivera et al. 2015). The stings of some *Conus* capable of subduing and killing a vertebrate prey or predator can be deadly to humans (Kohn 2018). The discovery of greatly diversified, short peptides in *Conus* venoms that predominantly target ion channels and receptors in the nervous system with unprecedented selectivity (Olivera et al. 1985; Terlau & Olivera 2004) has led to growing multidisciplinary interest in these venoms. To date, venom peptides from over 300 species of cone snails have been characterized using various approaches (Puillandre et al. 2016; Phuong et al. 2018, 2019), establishing *Conus* venoms as an excellent model system for studying molecular evolution in venomous marine animals. Furthermore, given their unique diversity and selectivity profile, cone snail venom peptides have become indispensable pharmacological probes and have been developed as drugs and drug leads for pain, epilepsy, diabetes and stroke (Prashanth et al. 2014, Safavi-Hemami et al 2019; Balsara et al. 2015; Jimenez et al. 2002; Robinson & Safavi-Hemami 2017).

While cone snail venoms have become a focus of intensive pharmacological studies, present knowledge of the venoms of conoideans outside the family Conidae has lagged far behind (Gorson et al. 2015; Puillandre et al. 2016). Given the vast diversity of over 10,000 estimated non-conid conoidean species (compared to ~900 species of Conidae) venom composition reports in the relatively well studied genus *Conus* may merely be scratching the surface of the vast diversity of venom compounds evolved in the superfamily Conoidea (Abdelkrim et al. 2018).

This is in part a result of methodological difficulties – most conoideans are small snails (Bouchet et al. 2009), with sometimes exceedingly small venom glands, which cannot yield a sufficient

amount of venom for biochemical characterization. However, the recent advent of deep sequencing not only provides a methodological basis for generating high quality data from miniscule amounts of venom, but also provides the opportunity to efficiently address questions of venom composition and evolution of venom gene families in a single study. Therefore, it is to be anticipated that our understanding of venom evolution across the conoidean tree of life will improve drastically in the upcoming years, allowing for the harnessing of the vast pharmacological potential of the superfamily Conoidea. The first transcriptomic analyses of the conoidean families Turridae, Pseudomelatomidae (Gonzales & Saloma 2014) and Terebridae (Gorson et al. 2015) gave a glimpse of the venom composition in the conoidean lineages divergent from Conidae, but currently available data remain too fragmentary to attempt comparative analysis across Conoidea.

In this study we present a comprehensive survey of the venom gland transcriptomes of two species of the genus *Clavus* Montfort, 1810, which belongs to the family Drilliidae Olsson, 1964, one of the major conoidean families whose venoms are uncharacterized. Its members retain a set of plesiomorphic characters in morphology, which led many authors to place Drilliidae at the base of conoidean radiation (Taylor et al. 1993). Nevertheless, with about 530 living species currently classified to it (MolluscaBase, <http://www.molluscabase.org/aphia.php?p=taxdetails&id=23032>), this family is among the highly diversified lineages of Conoidea as well (Abdelkrim 2018), (Fig. 1A). The unique structure of drilliid radulae implies a principally different mechanism of prey envenomation (Sysoev & Kantor 1989). Nevertheless, there is no published data on the diet and feeding biology of *Clavus* (and Drilliidae in general). Transcriptomic profiling reveals a vast diversity of the putative *Clavus* venom components that we term **drillipeptides**. Although biological activity of these components is still to be established, their number in the transcriptomic dataset notably exceeds the toxin diversity observed in individual specimens of *Conus* analyzed employing the similar methodology. The complement of putative toxins observed in individual specimens of *Clavus*, while closely parallel to *Conus* venoms in many respects, is clearly unique among Conoidea, most notably in the diversity of toxins found in each species analyzed.

Material and Methods

Specimen collection

All studied specimens were collected in the proximity of Olango Island (Central Philippines) on August 5, 2015. Specimens of *Clavus canalicularis* (Röding, 1798) and *Clavus davidgilmouri* [Fedosov & Puillandre in press] (referred to as CvCn1, CvCn2 for *Clavus canalicularis* specimen 1 and 2, respectively, and CvDg1 and CvDg2 for *Clavus davidgilmouri* specimen 1 and 2, respectively) were recovered from the same habitat – coral rubble at depths of ~15 – 25 meters by local fishermen and delivered to the laboratory alive on the same day, or the day after they were collected. After breaking the shell, the mantle was removed, and the body haemocoel dissected to extract the venom gland. Dissected venom glands were immediately suspended in *RNAlater*TM (Thermo Fisher Scientific). A tissue clip was taken from each dissected specimen and preserved in *RNAlater*TM for subsequent DNA extraction to confirm initial species identification. Samples in *RNAlater*TM were kept at room temperature for 24 hours, and then frozen and kept at -20°C until RNA extraction.

DNA extraction, amplification and phylogenetic analysis

To confirm identity and to ensure conspecificity of the specimens dissected for transcriptomic study, a barcode fragment of mitochondrial gene cytochrome oxidase I (COI) was sequenced for the four dissected specimens of *Clavus*, as well as for nine additional specimens. Total genomic DNA was extracted from foot muscle tissue using Thermo Scientific Gene-JET Genomic DNA Purification 50 Kit following the manufacturer's instructions. Fragments of the COI gene were amplified with the standard pair of primers LCO1490 and HCO2198 (Folmer et al. 1994). PCRs were performed in 20 µL final volume containing approximately 3 ng template DNA, 1.5 mM MgCl₂, 0.26 mM of each nucleotide, 0.3 µL of each primer, 5% DMSO and either 0.75 µL of Taq Polymerase (Qbiogene) or BioHYTaq DNA polymerase (Dialat, Moscow). The PCR profile for the COI started with 5 min at 95°C followed by 40 cycles with the denaturation at 95°C (35 s), annealing at 50°C (35 s) and elongation at 72°C (1 min), with final elongation phase at 72°C (7 min). The amplified DNA fragments were sequenced in both directions to confirm accuracy of each sequence. The sequencing was performed in the SIEE RAS molecular facility on an ABI 3500 Genetic analyser. Alternatively, the purified PCR products encoding *Clavus* COI sequences were annealed to a plasmid vector and transformed into competent cells as previously described (Fedosov et al., 2011). The nucleic acid sequences of six individual COI-encoding clones from each specimen were determined according to the standard protocol

for automated sequencing at the Health Sciences Center Core Sequencing Facility, University of Utah, and compared to ensure sequence consensus.

A total of 39 specimens were included in the phylogenetic dataset, including 3 species of the conoidean family Horaiclavidae as outgroups (taxonomy of the Conoidea after Bouchet et al. 2011). Of the 36 Drilliidae specimens analyzed, 17 belonged to the genus *Clavus*, including 13 specimens sequenced in the present study, and 4 records recovered from the GenBank sequencing database; all specimens originated from the Central Philippines. Sequences were aligned using ClustalW, as implemented in BioEdit 7.0.9.0 (Hall 1999). The 658 nucleotide positions in the COI alignment were treated as three independent partitions, corresponding to the three codon positions in the RaxML analysis (Stamatakis, 2006). GTR+I+G was estimated as best fitting nucleotide substitution model by Modelgenerator v.85 (Keane et al. 2006) under Akaike information criterion. Robustness of nodes was assessed using the Thorough Bootstrapping algorithm (Felsenstein, 1985) with 1,000 iterations. Phylogenetic analysis was performed on the Cipres Science Gateway (<http://www.phylo.org/portal2>), using RAXML-HPC2 on XSEDE (Miller et al. 2010). The matrix of pairwise genetic distances for the COI alignment was calculated using MEGA 6 (Tamura et al. 2013).

RNA extraction and sequencing

For transcriptome sequencing, total RNA extraction from venom glands was performed using the Direct-zol RNA extraction kit (Zymo Research, CA), with on-column DNase treatment, according to the manufacturer's instructions. Venom glands were homogenized in Direct-zol TRI reagent using the TissueRuptor II (Qiagen) with a disposable probes to avoid sample cross-contamination. cDNA library preparation and sequencing were performed by the University of Utah High Throughput Genomics Core Facility. Briefly, total RNA quality and quantity were first determined on an Agilent 2200 TapeStation (Agilent Technologies). A dual-indexed library was constructed with the Illumina TruSeq Stranded mRNA Sample Prep Kit with oligo (dT) selection and an average insert size of approximately 220 bp. The library was validated on an Agilent 2200 TapeStation and using a qPCR assay (Kapa Biosystems Library Quantification Kit for Illumina). 125 cycle paired-end (PE) sequencing was performed on an Illumina HiSeq2000 instrument (San Diego, CA). The statistics of the obtained raw reads data are summarized in Supplementary Table 1.

Transcriptome assembly

Each set of PE raw reads was first interleaved and processed by FQTRIM (version 0.9.4, <http://ccb.jhu.edu/software/fqtrim/>) and prinseq-lite (version 0.20.4) (Schmieder and Edwards 2011) to remove adapter and low quality base pairs or reads. Bases at the 3'-end were trimmed until the quality score was above 20, and polyA/T tails in good quality were kept. Reads with at least one base quality score below 10 or less than 80 bps in length were discarded. RNA reads remaining after quality control were error corrected (each set of reads separately), and then assembled using TRINITY v2.2.0 (Haas et al. 2013) with kmer size used to build the De Bruijn Graphs set to 31 (kmer size was optimized for *Conus* transcriptome assembly as described in Li et al., 2017). The assembly metrics for the four specimens of *Clavus* are summarized in the supplementary Table 2.

Transcript annotation

Transcriptome assemblies were annotated using NCBI-BLASTX against a combined ConoServer (Kaas, et al. 2012) and UniProtKB database (release April 2015, blastp e-value < **10e-10**). The conotoxin sequences published in 2015-2016 that were not yet uploaded into Uniprot at the time of the analysis (Barghi, et al. 2015a, b; Phuong, et al. 2016) were also added to our in-house database. Transcript contigs that did not have significant hits in known databases were translated into six frames, and predicted open reading frames (containing a start and stop codon) with a total length more than 50 amino acids (AAs) were retained for further analysis. Subsequently, the predicted sequences were analyzed, and only those with N-terminal signal peptides [predicted using SignalP (Petersen, et al. 2011), D value > 0.7] but no transmembrane topology [predicted using Phobius (Käll, et al. 2007)] were kept. Transcriptome assembly software programs are known to generate misassembled contigs, particularly for sequences with low read coverage. To address this we only report transcripts that were identified in at least two specimens and with tpm values >100. Prior *Conus* transcriptome studies have demonstrated that using assembly software alone (e.g., Trinity) does not reveal the full diversity of transcripts, particularly for those deriving from polymorphic loci (Phuong et al. 2016; Robinson et al. 2017; and Macrander et al. 2018). Remapping of trimmed reads onto assembled sequences can partially retrieve these otherwise evasive sequences (Robinson et al.

2017; and Macrander et al. 2018). However, this has to be done semi-manually and, in our hands, is too subjective to be reproducible. We therefore only report contigs assembled by Trinity but would like to emphasize that this methodology might not fully retrieve the genetic diversity of venom peptide sequences in *Clavus*.

Many transcriptomic analyses of the genus *Conus* assume varying degrees of signal sequence identity (Barghi et al. 2015a; Peng et al. 2016), which reflect *a priori* knowledge of signal sequence conservation among toxin gene superfamilies. The venom composition of the conoidean family Drilliidae, however, remains *terra incognita*, and it is not only that the number of toxin gene superfamilies expressed in the venom is unknown, but also that there is no information regarding the level of conservation of the signal sequences in relation to the mature peptide regions. This hampers inference of the common origin of the predicted venom peptides, which is necessary to delineate gene superfamilies. In particular, it is unclear whether the gene superfamilies in *Clavus* venom can be established based on comparable degrees of sequence identities as in *Conus*. To tackle this issue, we generated alternative clustering schemes based on different sequence identity thresholds, and then rated them to select an optimal scheme that would define putative gene superfamilies. To infer clusters of *Clavus* venom gland transcripts based on the similarity of their signal sequences (Puillandre et al. 2012) we used CD-Hit (Huang et al. 2010) with percent sequence identity threshold values of 75, 70 and 65. However, even the 65 identity threshold resulted in a very high number of putative gene superfamilies (>180), which would complicate comparisons with conotoxins. Therefore, we further reduced the percent signal sequence identity values to 60, 55 and 51%.

The drawback of the CD-Hit is that there is no clear measure to compare alternative cluster sets and select the one that best fits the dataset. To rate the six obtained clustering schemes, we used two approaches. First, we estimated the level of structural heterogeneity of the mature peptide region based on the number of Cys-patterns in the inferred clusters (as the cysteine framework is an important determinant of the physiological activity of *Conus* toxins and is used for toxin classification – Terlau & Olivera 2004). Second, we classified inferred transcripts based on 1) BLAST hits of the entire precursor against the external databases (i.e. *Clavus-Clavus* hits, and *no blastx hit* entries were removed from the dataset) and 2) sets of orthogroups recovered for the entire precursor by OrthoFinder 2.1.2. (also see below). The 105 BLAST hits were classified in 23 clusters, and 206 orthogroups with 2 or more sequences in each were returned by Orthofinder. Subsequently, an in-house Python script that implements a pair-counting approach (Wagner & Wagner 2007) was used to estimate how well each of the proposed signal

sequence-based cluster sets matches the BLAST-based cluster set. As the identified orthogroups are certainly more conserved, another Python script has been used to calculate how many of the identified orthogroups were split between two or more putative gene superfamilies. These numbers were used as penalty scores: the lower the score, the closer the match between the signal sequence-based clusters and entire precursor-based orthogroups. The signal sequence-based set of clusters that returned the highest score in comparison with BLAST clusters, and lowest penalty in comparison with orthogroups was selected for putative gene superfamily breakdown.

All-by-all NCBI-BLASTN searches were carried out to identify reciprocal best-hit (rbh) pairs among the inter/intra-specific drillipeptide nucleotide sequences. ClustalW alignment was then used to globally align the rbh pairs (Larkin, et al. 2007). The alignment results were parsed to obtain the percent identity of each aligned rbh pair. To investigate interspecific differences at the sequence level we performed comparative sequence analyses between the putative venom repertoires of each species (combined *Cv. canalicularis* dataset versus combined *Cv. davidgilmouri* dataset). All-by-all NCBI-BLASTN searches identified 177 and 165 rbh pairs between the two specimens of each *Clavus* species (*Cv. canalicularis* and *Cv. davidgilmouri*, respectively) and 209 rbh pairs when comparing the two species to one another (Fig. 3).

To detect ortho-groups in the *Cv. canalicularis* and *Cv. davidgilmouri* transcriptomes, we used OrthoFinder 1.0.6 (Emms & Kelly, 2015) with the two input files corresponding to combined *Cv. canalicularis* and combined *Cv. davidgilmouri* datasets, with default settings. Separately for the L-Rhamnose binding lectins and insulins, the low TPM transcripts were also recovered from the assembly and added to the dataset. To calculate correlation between the *Clavus* species in transcript diversity and expression levels, Pearson's R was calculated using the numpy method (numpy.corrcoef).

Structure prediction of toxin precursors

Conotoxin precursors are typically composed of an N-terminal signal peptide, a mature peptide region and an intervening propetide region between the signal peptide and mature peptide (e.g. Terlau & Olivera 2004). Some precursors also contain a postpeptide C-terminally of the mature peptide. For the *Clavus* dataset, the signal peptide was characterized using SignalP 4.0 server (Petersen et al., 2011). Although the exact boundary of the mature peptide is difficult to assign

without peptide level confirmation due to possible alternative cleavage events, (Lu et al. 2014) , a canonical mature peptide region was still predicted for each transcript in order to analyze *drillipeptide* precursor structure. The *Clavus* cysteine-rich peptide precursors were first predicted by a custom PERL script and then confirmed using Conoserver (Kaas et al., 2012) and ProP 1.0 server (Dukert et al., 2004). In peptide precursors containing no cysteine residues, the mature peptide regions were mainly predicted according to sequence alignment to known similar sequences. The custom PERL script to facilitate canonical mature peptide identification was designed to predict cleavage sites based on the published data (Fan & Nagle, 1996; Fan et al. 2000; Dukert et al., 2004; Bousquet-Moore et al. 2010; Kaas et al. 2012). Briefly, the N-terminal boundary of the mature peptide was first identified C-terminally of two consecutive basic amino acids (RR-, KR-, RK-, KK-) before and closest to the first cysteine residue after removing the signal peptide from the precursor. If a single basic amino acid remained between the N-terminus and the first cysteine, and the distance to N-terminus exceeded 5 AAs, C-terminal of this basic amino acid was considered as the N- terminal boundary of the mature peptide. The C-terminal boundary of the mature peptide was identified by the presence of peptidylglycine α -amidating monooxygenase (PAM) enzyme recognition sites consisting of a glycine followed by a basic or dibasic residue (-GKR-, -GRR-, -GRK- -GK-, -GR-, -KR-, -RR-) at the C-terminal part of the last cysteine residue.

Results and Discussion

Phylogenetic analysis identifies two studied species as close relatives

The phylogenetic analysis based on the barcode fragment of COI placed the dissected *Clavus* specimens in two different well supported clades, corresponding to *Cv. canalicularis* and *Cv. davidgilmouri* (Fig. 1B). While the *Cv. davidgilmouri* clade appeared to be homogeneous (highest within-clade K2P of 0.7%), the *Cv. canalicularis* clade showed clear genetic structure, with the highest within-clade genetic distance of 3.4%. The two *Cv. canalicularis* specimens studied for the venom gland transcriptome, although collected from same locality, showed a genetic distance of 2.7%, which is notably higher than between the two studied specimens of *Cv. davidgilmouri*. Nevertheless, the inferred pairwise comparisons within each clade fall within a range of intraspecific genetic distances demonstrated for the superfamily Conoidea (Puillandre et al. 2010; Fedosov, Puillandre 2012; Fedosov et al. 2017). The mean between-clades K2P

genetic distance is 8.1%, and smallest – 6.9%, which indicates that the two species, although closely related, are not sister species consistent with the previous results of Fedosov et al. (in preparation).

Clavus putative toxin diversity is unparalleled among Conoidea

A total of 1,176 unique putative toxin-like peptides were identified in the combined *Clavus* dataset. They constitute 1.85% of the number of unique contigs recovered, and account for 28.8% of the total gene expression level (calculated from tpm values).

Typically, the identified transcripts demonstrated a canonical organization with a 16 – 34 AA signal sequence, followed by a pro-region of varying length (2 to 263 AA residue, lacking in ~28% of identified putative precursors), a mature peptide region and, in ~30% of identified transcripts, a short, one or few AA post region.

The number of clusters recovered by CD-Hit for the signal sequences of the putative *Clavus* toxins ranged from 120 (sequence identity threshold 51%) to 244 (sequence identity threshold 75%). In all clustering schemes, the number of Cys-patterns per cluster ranged from one to nine, and the proportion of clusters with only one Cys-pattern ranged from 0.55 to 0.82 when the sequence identity threshold was set at 51 and at 75 percent, respectively (Supplementary Figure 2). The CD-Hit clusters recovered when the signal sequence identity was set at 60 most closely match the clusters inferred by BLAST and Orthofinder (Supplementary Table 3), and therefore, they were used to designate the putative *Clavus* venom gene superfamilies. The resulting partitioning scheme includes as many as 158 putative gene superfamilies, each comprising one to 110 predicted toxins. At present, we propose naming them by prepending a superfamily number by the term “*Drillipeptide*”, however, more elaborated or convenient nomenclature may be developed in the future, as our knowledge of drillipeptides grows. Members of only 39 gene superfamilies had external BLAST Hits. From here onwards we will refer to the venom peptides that demonstrate sequence similarity to known groups of *Conus* toxins as A-like drillipeptides, F-like drillipeptide, etc. or drilliporins, drillinsulins and (drillidipines, drillikunitzins). However, we emphasize that this often remote sequence similarity does not necessarily imply sequence homology.

Among the putative toxins with BLAST hits, the 'MWIRK'-superfamily-like transcripts showed highest diversity and total expression level. Interestingly, each of the three highly expressed and diverse groups of peptides in *Clavus* venom, the 'MWIRK'-superfamily-like transcripts, drilliporins and L-Rhamnose-binding lectins, were represented by several gene superfamilies as defined by the signal sequence (four, three and four respectively). Conversely, only one drillipeptide gene superfamily produced BLAST hits to each of conotoxin superfamilies A, F, I3, M, O1, O2, U, *Conus* insulins, conodipine and konkunitzin.

The high diversity and expression levels of the 'MWIRK'-superfamily-like transcripts, drilliporins and L-Rhamnose-binding lectins in the *Clavus* venom gland transcriptomes suggest that they play an important role in prey capture or defense (see below). Nevertheless, additional data on the spectrum of biological interaction of *Clavus* spp. is needed to pose these data in ecological context. The fact that in some cases, transcripts with similar BLAST hits represent different gene superfamilies may be explained by either: 1) the signal sequence being less highly conserved compared to typical conoidean gene superfamilies, or 2) putative toxins with similar mature peptides having been recruited from different genes, or finally, 3) the reference sequences of conoporins and L-Rhamnose-binding lectins actually representing heterogeneous groups themselves.

The Shannon's index for transcript diversity was slightly higher in specimens of *Cv. davidgilmouri* (4.47 and 4.32) compared to *Cv. canalicularis* (4.24 and 4.26) (Supplementary Table 6). A recent transcriptomic study on two closely related species of *Conus*, *Conus tribblei* and *C. lenavati* (Barghi et al. 2015b), reported the highest toxin diversity of any cone snails characterized to date (Shannon's index 3.3 in both species). Shannon's index values in both species of *Clavus* exceed those in *C. tribblei* and *C. lenavati*, whereas the Shannon's evenness values were comparable and ranged from 0.9 to 0.93 and from 0.89 to 0.92 in the *C. tribblei*-*C. lenavati* and among *Clavus* species, respectively. While we believe that the vast majority of transcripts identified in this study are likely to be part of the toxin repertoire of *Clavus* we cannot rule out that a small subset of transcripts is not used in predation but as endogenous signaling peptides or falsely predicted to be translated into a secreted peptide product. Future proteomic studies are necessary to verify predicted venom peptides of *Clavus*, while functional evaluations will be necessary to unambiguously distinguish toxins from endogenous signaling peptides.

Putative Clavus venom repertoires are conserved at gene superfamily level, and highly divergent at mature peptide

644 and 532 unique transcripts were identified in the transcriptome of the venom glands of *Cv. canalicularis* and *Cv. davidgilmouri*, respectively. Comparative analysis of the putative toxin sequences between the two *Cv. canalicularis* specimens revealed that 89.3% of rbh pairs were greater than 95% identical, and of these, 65.0% of sequences were greater than 99% identical (Fig. 2A). In the two specimens of *Cv. davidgilmouri* 89.7% of rbh pairs were greater than 95% identical, and of these, 64.2% of sequences were greater than 99% identical (Fig. 2B). Finally, the comparative analysis of the combined *Cv. canalicularis* and *Cv. davidgilmouri* datasets revealed that 55.5% of putative toxin precursors shared >95% identity, but of these only 6.2% of sequences were more than 99% identical (Fig. 2C). These data indicate that despite high overall similarity between the venom gland transcriptomes of the two *Clavus* species, considerable variation exists at the individual sequence level.

The most diverse gene superfamilies (7, 93, 40, 76 and 77) were, represented in each *Clavus* species by 13 or more unique putative toxins and account for almost 25% of the total diversity of the identified transcripts (Fig. 2D). Overall, a moderately strong correlation in the number of unique sequences per gene superfamily was observed both between the two species of *Clavus* (Pearson's R, 0.83), and between the two studied specimens of each species (Pearson's R 0.89 and 0.85 for the *Cv. canalicularis* and *Cv. davidgilmouri*, respectively) (Fig. 2E, Supplementary Table 5). The slightly weaker correlation between the two specimens of *Cv. davidgilmouri* is probably a result of different assembly statistics, with twice as many contigs identified in CvDg2 compared to CvDg1. Conversely, when expression levels of the gene superfamilies are compared (Fig. 2F, G), considerable interspecific variation can be noted in most gene superfamilies. Furthermore, the gene superfamily expression levels vary greatly among the specimens of *Cv. canalicularis* compared to specimens of *Cv. davidgilmouri* (Pearson's R 0.62 and 0.85 for the *Cv. canalicularis* and *Cv. davidgilmouri*, respectively). The latter result cannot be attributed to artifacts of assembly (the assembly statistics are very similar between the two *Cv. canalicularis* specimens whose transcriptomes were analyzed). It can be, however, explained by the notably higher genetic divergence between the two *Cv. canalicularis* specimens, inferred from the COI sequences, suggesting that the profile of the venom gene family expression is highly volatile, and may vary considerably among conspecific individuals.

Nevertheless, the results of the RBH analysis in *Clavus* suggest that the repertoire of putative toxin gene families expressed in venom gland remains largely conserved within species, and shows divergence between the closely-related, (but not sister) species. This result closely parallels intra- and interspecific comparisons reported in two closely related sister species of *Conus*; *Conus praecellens* and *Conus andremenezi* (Li et al. 2017) that showed a similar degree of genetic differentiation compared to the *Clavus* species studied herein. The generally strong correlation in the number of unique transcripts per gene superfamily between *Cv. canalicularis* and *Cv. davidgilmouri* corroborates this pattern, implying similar degrees of molecular diversity across the main putative venom components in the two species. Finally, the observed high variation in the putative venom gene families in *Cv. canalicularis* may be explained by the notable genetic divergence between the two dissected specimens (Fig. 1B). Although a more rigorous estimate of the venom variation in *Clavus* is needed to validate and more accurately quantify the observed variation and disentangle its correlates, our results are consistent with gene superfamily expression being the most volatile parameter of venom composition. Being fine-tuned to niche parameters, venom gene expression may profoundly change starting at early stages of the speciation process. Both the profile of molecular diversity (assessed through the number of unique gene products per venom gene superfamily), and the venom composition at the sequence level, appear to be more conserved, and are expected to contribute to venom variation at a higher taxonomic level.

Clavus venom peptides are comparable to conotoxins in sequence length and structural properties

The violin plot (Fig. 3A) shows distribution of the mature peptide region lengths in the two species of the genus *Clavus* in comparison with published data on some *Conus* species. The pattern of mature peptide region lengths distribution is apparently very close between the two species of *Clavus*. The median length of the mature peptide sequence is at 41 AA for *Cv. canalicularis* and at 44 AA for *Cv. davidgilmouri*, and the lower quartile is at 32 AA in both species. 108 and 58 unique transcripts in the venom gland of *Cv. canalicularis* and *Cv. davidgilmouri* respectively show a mature peptide length of less than 25 AA residues. The putative toxins of the 20 gene families with mature toxin regions containing on average 13 – 30 AA residues closely match in this respect typical conotoxins (Figs. 3B, 4B, Supp. Fig. 2). The drillipeptides in the highly diversified superfamily 4, which shows the highest expression level

among identified families, are a perfect example of such peptides. For this gene family, a total of 64 individual peptides were predicted from our transcriptomic data (Fig. 4A), which formed 9 orthogroups as inferred from Orthofinder (Supplementary Figure 3). The mature peptide sequences range from 29 to 42 AA residues in length, and contain 6 Cys residues organized in the cysteine framework VI / VII (Fig. 4B).

The mature toxin regions are on average longer in other lineages of the 'Turrid' major clade of conoideans (Puillandre et al. 2011; Puillandre et al. 2016): *Gemmula speciosa* and *Unedogemmula bisaya* (Turridae) and *Crassispira cerithina* (Pseudomelatomidae) (Gonzales & Saloma 2014). The very limited transcriptomic data on *G. speciosa*, *U. bisaya*, and *C. cerithina* make it impossible, however, to test the significance of the difference in mature toxin lengths between Turridae and *Clavus* spp. The broad length range in drillipeptides with typical toxin precursor structure resembles transcriptomes of the Turridae, Pseudomelatomidae and Terebridae (Imperial et al. 2007; Gorson et al. 2015), whereas the predicted high diversity of relatively short, disulfide-rich peptides in the venom gland transcriptomes of *Clavus* spp. makes them closely comparable with *Conus* venoms. These features of *Clavus* venom gland peptides (high diversity, small size and predicted stability) suggest that drillipeptides, like conotoxins, may be promising candidates for future pharmacological investigations. Venom peptides with high expression levels can readily be explored by mass spectrometric analysis of *Clavus* venom, followed by chemical synthesis and activity testing. One such bioactive peptide, cdg14a, was identified by mass spectrometry and has been purified from the venom of *Clavus davidgilmouri*. This peptide elicits scratching and hyperactivity followed by a paw-thumping phenotype in mice (Chua et al., submitted). The precursor encoding cdg14a, CvDg14.4, from drillipeptide superfamily 7, was found in the transcriptomes of both *Clavus davidgilmouri* specimens (n1-815 tpm, n2-1,642 tpm). Figure 3C shows 23 and 22 venom gene superfamilies of *Clavus canalicularis* and *Cv. davidgilmouri* respectively that are both highly expressed (total tpm $\geq 1,000$) and comprise transcripts with mature peptide regions of ≤ 50 AA (area shown in gray). They could be first analyzed in a search of the most relevant target for subsequent structural studies and bioactivity assays.

While the majority of predicted peptides in *Clavus* venom share the same canonical structure of signal sequence, propeptide region followed by a single mature toxin with most conotoxin precursors, drillipeptide sequences are quite divergent. We therefore avoid inference of sequence homology between drillipeptides and conotoxins. Nevertheless, as with conotoxins, drillipeptides tend to have an even number of cysteines (typically six) in the mature peptide

region. Two of the most common canonical cysteine frameworks for conotoxins, VI/VII (C-C-CC-C-C) and I (CC-C-C), are well represented in the mature peptide regions of drillipeptides (Supplementary Table 8). However, statistical analysis of the most abundant cysteine frameworks in every category of sequences with an even number of cysteines (4-12) reveals that vicinal cysteines that are common in conotoxin cys-frameworks are not as frequent as scattered single cysteines in drillipeptides (Supplementary Table 8). Furthermore, the cysteine frameworks V (CC-CC) and III (CC-C-C-CC), each bearing two vicinal cysteines, were only identified in 0 and 2 drillipeptides, respectively despite being ubiquitous in conotoxins.

Sequence variability of Clavus insulins suggests a role in predation

Insulin and related peptides (insulin-like growth factors, relaxins and insulin-like peptides) are peptide hormones that are found throughout the animal kingdom where they play a role in regulating carbohydrate and fat metabolism and as neuromodulators of energy homeostasis and cognition (Blumenthal 2010; Gerozissis 2003; Smit et al. 1998). Insulin is synthesized as a precursor comprising three regions (A and B chains and a C peptide region), from which proteolytic cleavage of the C peptide in the Golgi releases the insulin heterodimer with A and B chains connected by disulfide bonds. The primary sequence and arrangement of disulfide-forming cysteines are highly conserved in all vertebrates, but invertebrate insulin family members are more variable. Molluscan and most other invertebrate insulins differ from vertebrate insulin in containing two additional cysteines that are assumed to form an additional disulfide between the A and B chain. Moreover, the mature molluscan insulin chains are typically larger than vertebrate insulins. We recently discovered that some species of cone snail weaponized their endogenous insulin for prey capture to induce insulin shock (dangerously low blood glucose levels) in prey (Safavi-Hemami et al. 2015, 2016; Ahorukomeye et al. 2019). *Conus* venom insulins are greatly diversified, specifically expressed in the venom gland, exhibit distinct sequence characteristics from molluscan signaling insulins and correlate with prey taxa (Safavi-Hemami et al. 2016). Cone snail species that prey on molluscs and polychaetes express insulins that closely resemble invertebrate insulins, while a subset of fish-hunting species expresses insulins with high sequence and structural similarity to fish insulins which are capable of activating the fish and human insulin receptor (Safavi-Hemami et al. 2015, 2016; Ahorukomeye et al. 2019).

Here, we identify seven putative venom insulins in *Clavus* that share several characteristics with *Conus* venom insulins, including significant sequence diversification and high expression levels of some sequences. *Clavus* venom insulins are derived from a single gene superfamily as evidenced by their conserved signal sequence (Fig. 5A) but exhibit notable sequence divergence within the A and B chains, a hallmark of *Conus* venom insulins (Fig. 5A). This includes high diversification rates within the primary amino acid sequences of the A and B chains (apart from a conserved cysteine framework) and changes in chain lengths, most likely resulting from insertion and deletion events (Fig. 5A). This significant sequence divergence of *Clavus* insulins combined with the very high expression level of drillinsulin CvDg1 in *Clavus davidgilmouri* (n1 - tpm 12,576, n2 – tpm 3,084), suggests a role of *Clavus* insulins in prey capture. Interestingly, insulins sequenced from *Clavus* share structural similarity with worm and mollusk insulins suggesting that *Clavus* specializes on invertebrate prey as previously reported for snail and worm-hunting *Conus* (Safavi-Hemami et al. 2016) (Fig. 5B). This is in stark contrast to insulins sequenced from fish-hunting cone snails that exhibit the vertebrate cysteine framework (Fig. 5B). A more accurate interpretation of these observed patterns will require additional studies to shed light on structure-function properties of the mature venom insulins, their sequence comparison with *Clavus* signaling insulins and on the spectrum of *Clavus* prey taxa.

Phylogenetic analysis classifies the seven drillinsulins into 3 clusters (Fig. 5C left). The cluster 1 drillinsulin CvDg1 shows very high expression levels in *Cv. davidgilmouri* (n1 - tpm 12,576, n2 – tpm 3,084) and was also detected in *Cv. canalicularis* n2 (drillinsulin CvCn4 - tpm 435). Conversely, the cluster 3 insulins (drillinsulin CvCn1 - tpm 865, drillinsulin CvCn3 - tpm 372) have higher expression in *Cv. canalicularis*, and the cluster 2 insulin is only found in *Cv. canalicularis* (drillinsulin CvCn2 - tpm 555, drillinsulin CvCn2ii - tpm 383). Two insulins, drillinsulin CvCn2 (tpm 555) and drillinsulin CvCn2ii (tpm 383) from CvCn1 and CvCn2 respectively have one amino acid substitution, and likely represent allelic variation of the same locus. In the reconstructed phylogeny of conoidean venom insulin precursors (Figure 5C right), the *Clavus* insulins cluster with the insulins of the Turridae species (PP 0.93), and the *Conus* venom insulins form a separate, well supported cluster (PP 1), which reflects phylogenetic relationships among Drilliidae, Turridae and Conidae (Abdelkrim et al. 2018). In contrast, the phylogenetic analysis of the mature peptide sequence (Figure 5C left) does not resolve the relationships among the venom insulins of *Clavus*, Turridae and *Conus*, although it recovers them in three reciprocally monophyletic clades.

Molecular diversity of porins in Clavus venom

Porins were among the most diversified and highly expressed venom components in all four profiled specimens of *Clavus*. The 25 detected porin transcripts range from 174 to 247 AA in length and share a stretch of about 70 AA in the mature peptide region with high similarity across all sequences (Supp. Fig. 4). Nevertheless, these sequences form three clusters distinctive by signal sequence (referred to three separate gene superfamilies, MALLR, MALTL, and MKLFL/MKMFV). Among them, the MALLR sequences show identity to coluporin-12 and coluporin-21 of *Cumia reticulata*. The divergent predicted peptide, here referred to as claviporin-Cn6, shows 32% sequence similarity to tereporin-Ca1 from *Terebra anilis*, and predicted porins in the MALTL gene superfamily display 36-40% sequence similarity to the conoporins from *Conus lividus* and *C. ebraeus*. Therefore, here we consider all three clusters to be porins based on the demonstrated similarity in the mature peptide region (Supp. Fig. 4), which also suggests conserved secondary structure. Porins were represented by four to ten unique transcripts per specimen and were typically highly expressed, with a total expression ranging from tpm 4,682 (CvCn1) to tpm 15,072 (CvDg1).

Porins are a class of peptides that bind to, and form pores in cellular membranes (Iacovahe et al. 2010). Because of their cytolytic effect, porins are involved in host-pathogen interactions, as the agents facilitating infection by a pathogen, as well as a component of the immune response of the host. Porins are a common component of animal venoms, wherein, by disrupting cellular membranes they are thought to facilitate spread of the venom in prey tissues (Parker and Feil 2005). Among marine invertebrates the best studied are the porins of sea anemones (actinoporins) (Alvarez et al. 2009; Rivera-de-Torre et al. 2016; Podobnik and Anderluh 2017), and in gastropods - the echotoxins of tonnoideans (Kawashima et al. 2003; Bose et al. 2017) and coluporins of the Mediterranean bloodsucking neogastropod *Cumia reticulata* (Gerdol et al. 2018). In *Cumia*, documented notable expansion of the porin gene superfamily is explained by their functional importance for penetrating prey circulatory system, and by their haemolytic effect. Despite Gerdol and coauthors' suggestion (2018) that porins are only present in two conoidean lineages, cone-snails (in particular *Conus geographus* - Safavi-Hemami et al. 2014), and the family Terebridae (Gorson et al. 2015), our results demonstrate that porins are not only present, but are also highly diversified and expressed in *Clavus*, and, possibly, in other drilliids.

Furthermore, the phylogenetic analysis of the *Clavus* porins with other neogastropod porins (Gerdol et al. 2018) revealed that they actually fall into two separate clusters (Fig. 6). The MALTL superfamily groups with the conoporins, whereas the remaining two clusters form a clade nested in the coluporins, and both relationships are highly supported. This result has several important implications. Porins of *Cumia* were shown to be divergent from all previously known porins of Conoidea (Gerdol et al. 2018), mirroring distant relationship between Conoidea and Buccinoidea. Such concordance of the species tree and the gene tree recovered for one or another gene superfamily characterized from transcriptomic data is a predominant pattern in studies of gastropod venoms (for example, Terebridae - Gorson et al. 2015; *Cumia* - Gerdol et al. 2018; *Profundiconus* - Fassio et al. 2019). The contrasting example is the relationships of *Conus* insulins, which reflects the relationship among the prey taxa (Safavi-Hemami et al. 2016). The phylogeny of *Clavus* porins, however, shows a third pattern, which is different from the two mentioned above. It is consistent with porins having been recruited from two separate gene superfamilies: one presumably was present in the common ancestor of Conoidea and *Cumia* (but never previously detected in Conoidea), and the other one, on the contrary, only currently known in Conoidea. The presence of products of both these gene clusters in *Clavus* may suggest that their molecular targets (e.g. in vertebrate vs invertebrate systems) and/or action mechanisms differ from one another. This hypothesis is consistent, for example, with the distinction between the defensive and predatory porins. Further studies on the diversity of porins in Neogastropoda, and on the spatial differentiation of gene expression in the *Clavus* venom gland, and the spectrum of molecular activities are necessary to better understand diversity and evolution of venom porins in the Neogastropoda.

Structural diversity of Rhamnose-binding lectins

Carbohydrate-binding proteins, lectins, are ubiquitous and diverse in nature. Rhamnose-binding lectins (RBLs), which play roles in innate immunity in fish (Watanabe et al, 2009), formed one of the remarkably diverse clusters of the predicted venom components of *Clavus* spp. Six predicted *Clavus* venom transcripts with expression levels above 100 tpm showed detectable similarity to the L- RBLs of the Chum salmon *Oncorhynchus keta* and the Amur catfish *Silurus asotus*. When the transcripts with lower expression levels were considered, a total of 28 transcripts encoding RBLs were identified, their tpm expression levels ranging from 5 to 15,382. The majority (24 transcripts) of the predicted RBL precursors comprised 2 carbohydrate

recognition domains, CRDs (as defined by Ogawa et al. 2011), in which, it is interesting that the two CRDs of the same sequence have the same number of cysteines. The alignment of the CRDs is provided in Supplementary Figure 5 with the secondary structure predictions mainly based on the annotation provided by Ogawa and co-authors (2011). Although the CRDs from Drillilectin_CvCn1, the most abundant RBL from *Cv. canalicularis* (indicated in Supplementary Figure 5 by green arrows, tpm: 1542.16) has the canonical eight cysteines to form four disulfide bonds, the CRDs from Drillilectin_CvDg1, the most abundant RBL from *Cv. davidgilmouri* (indicated with red arrows, tpm: 15382.68) has six cysteines and a short sequence lacking the C1-C3 disulfide bond. In addition, one predicted RBL transcript (Drillilectin_CvCn11, highlighted with orange in Supplementary Figure 5) comprised 3 CRDs and three transcripts (Drillilectin_CvDg3, Drillilectin_CvCn5, and Drillilectin_CvDg10, highlighted with purple) comprised 4 CRDs. All CRDs in Drillilectin_CvCn11, Drillilectin_CvCn5 and Drillilectin_CvDg3 have only three disulfide bonds, but all CRDs in Drillilectin_CvDg10 have four. The CRDs in *Clavus* RBLs typically contain the 'YGR' and 'DPC' motifs characteristic for other RBLs identified from marine animals (e.g. sea anemone, urchin, oyster or fish), despite the overall sequence identity of the *Clavus* RBLs to previously identified RBLs being rather low (on average $\leq 30\%$). The high diversity and, in some cases very high expression levels of lectins suggest that they may play an important role in envenomation. This hypothesis for Conoidea in general is supported by the evidence of high RBL differential expression in the venom gland of *Profundiconus vaubani*, the most early divergent cone-snail genus, and a distant conoidean relative of *Clavus* (Fassio et al. 2019).

Both lectins and porins bind to the components of cellular membranes, and both are involved in immunity, due to their demonstrated ability to disrupt cellular membranes (Parker and Feil 2005; Watanabe et al, 2009). The high diversity and expression levels of both these protein classes in the venom of *Clavus* can possibly be explained by their action, which may facilitate venom penetration and spreading in the prey body (consistent with the hypothesis of Gerdol et al on *Cumia*). The radula of *Clavus*, with its flat marginals and comb-like lateral teeth, cannot be used to inject venom into the prey's body, as in *Conus*. However, based on the anatomical studies, drilliids use marginal teeth in a similar manner - holding a tooth at the tip of the proboscis presumably to make lacerations on the prey's integument (Sysoev & Kantor, 1989). Through the laceration venom can penetrate the prey's body, but to cause major effects on the prey's physiological circuits, it should reach and be spread by the circulatory system. This local diffusion of venom can be faster if adjacent epithelia are damaged by, e.g., cytolytic

polypeptides. Thus the high abundance and diversity of lectins and porins - proteins with demonstrated cytolytic activity - could be a molecular adaptation of *Clavus* in the lack of efficient mechanical venom delivery achieved in *Conus* by means of hypodermic radular teeth.

Conclusions

The venom gland transcriptomes of previously unstudied predatory gastropods of the genus *Clavus* demonstrates unique biochemical features: displaying a diversity of gene families unparalleled among conoideans, most of which do not show similarity to known toxin families. Despite the phylogenetic position of *Clavus* in the ‘turrid’ major clade of the Conoidea (that diverged from the ‘conid’ major clade, including the genus *Conus* around 130 Mya - Abdelkrim et al. 2018), some putative *Clavus* venom peptides show detectable similarity to conotoxins. Furthermore, our data suggest that the venoms of *Clavus* are dominated by rather short peptides, and in this regard resemble more the venom of *Conus* than venoms of more closely related Turridae and Terebridae. The demonstrated high diversity of short cysteine-rich peptides encoded in the venom gland transcriptomes of *Clavus* suggests they may also represent a valuable resource for drug discovery. Future studies on the characterization of the physiological effects and the identification of molecular targets of these peptides will elucidate the pharmacological potential of drillipeptides. Alongside the short cysteine-rich peptides, the venom of *Clavus* is remarkable for the high diversity and expression levels of insulins, rhamnose-binding lectins and porins. Insulins are likely to disrupt energy metabolism of prey while porins and lectins may exert cytolytic effects that enhance the spread of venom in the prey’s body. This is the first convincing evidence of complex biochemical adaptation leveraging the lack of efficient venom delivery apparatus in the ‘turrid’ lineage of Conoidea. Furthermore, it is certain that venom insulins, lectins and two gene families of porins were already present in the common ancestor of *Conus* and *Clavus* – i.e. in one of the earliest conoideans. Conversely, the available data is not sufficient to infer whether some known conotoxin gene superfamilies have homologs in the transcriptome of *Clavus*. Thus we reveal a novel, and in many aspects, a unique fragment of knowledge of conoidean venom evolution, which will likely inspire future studies on this remarkable, yet still poorly studied taxon.

Acknowledgments

The specimens used in this study were obtained in conjunction with a collection trip supported in part by the ICBG grant #1U01TW008163 and by the DOST PCHRD DDHP2 project (to G.P.C.).

We would like to thank Noel Saguil for help with sample collection, which was performed under the Department of Agriculture (DA) – Bureau of Fisheries and Aquatic Resource (BFAR)-issued gratuitous permit no. 0111-16. We thank Dr. Yuri Kantor for sharing photos of the radula of *Clavus canalicularis*, and David Massemin for kindly providing us photos of life specimens of *Clavus*.

This work was supported in part by National Institutes of Health Grants GM 48677 (to B.M.O), the Villum Young Investigator Grant (19063 to H.S.-H.), National Natural Science Foundation of China (Grant No.31500626, to A.L), Shanghai Natural Science Foundation of China (Grant No.19ZR1461600 to A.L.) and the Russian Science Foundation (Grant 19-74-10020 to A.F.). Research reported in this publication utilized the High-Throughput Genomics and Bioinformatic Analysis Shared Resource at Huntsman Cancer Institute at the University of Utah and was supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA042014.

Data accessibility

Raw sequence data analyzed in the present paper can be found at NCBI PRJNA610292.

References

- Abdelkrim J, Aznar-Cormano L, Fedosov AE, Kantor Y, Lozouet P, Phuong M, et al. (2018) Exon-capture based phylogeny and diversification of the venomous gastropods (Neogastropoda, Conoidea). *Molecular Biology and Evolution*, 35: 2355–2374.
- Ahorukomeye P, Disotuar MM, Gajewiak J, Karanth S, Watkins M, Robinson SD, Flórez Salcedo P, Smith NA, Smith BJ, Schlegel A, Forbes BE, Olivera B, Hung-Chieh Chou D, Safavi-Hemami H (2019) Fish-hunting cone snail venoms are a rich source of minimized ligands of the vertebrate insulin receptor. e41574.
- Álvarez C, Mancheno JM, Martinez D, Tejuca M, Pazos F, Lanio ME. 2009. Sticholysins, two pore-forming toxins produced by the Caribbean Sea anemone *Stichodactyla helianthus*: their interaction with membranes. *Toxicon* 54(8):1135–1147.

- Balsara R, Dang A, Donahue DL, Snow T & FJ, Castellino (2015) Conantokin-G attenuates detrimental effects of NMDAR hyperactivity in an ischemic rat model of stroke. *Plos ONE*, e0122840.
- Barghi N, Concepcion GP, Olivera BM & AO Lluisma (2015a) High conopeptide diversity in *Conus tribblei* revealed through analysis of venom duct transcriptome using two high-throughput sequencing platforms. *Marine Biology*, 17: 81-98.
- Barghi N, Concepcion GP, Olivera BM & AO Lluisma (2015b) Comparison of the venom peptides and their expression in closely related *Conus* species: Insights into adaptive post-speciation evolution of *Conus* exogenomes. *Genome Biology and Evolution*, 7: 1797-1814.
- Bose U, Wang T, Zhao M, Motti CA, Hall MR & SF Cummins (2017) Multiomics analysis of the giant triton snail salivary gland, a crown-of-thorns starfish predator. *Scientific reports*: 7: 1-14.
- Bouchet P, Kantor YI, Sysoev AV & N Puillandre (2011) A new operational classification of the Conoidea (Gastropoda). *Journal of Molluscan Studies*, 77: 273-308.
- Bouchet P, Lozouet P & AV Sysoev (2009). An inordinate fondness for turrids. *Deep-Sea research II* 56, 1724-1731.
- Blumenthal S. (2010) From insulin and insulin-like activity to the insulin superfamily of growth-promoting peptides: a 20th-century odyssey. *Perspect Biol Med* 53(4):491-508.
- Chua, V.M., Gajewiak, J., Watkins, M., Espino, S.S., Ramiro, I.B.L., Omega, C.A., Imperial, J.S., Carpio, L.P.D., Fedosov, A., Safavi-Hemami, H., Salvador-Reyes, L.A., Olivera, B.M., Concepcion, G.P. Purification and Characterization of the Pink-Floyd Drillipeptide, a Bioactive Venom Peptide from *Clavus davidgilmouri* (Family Drilliidae). *Biochemistry*. submitted.
- Duckert P, Brunak S, Blom N. 2004. Prediction of proprotein convertase cleavage sites. *Protein Engineering, Design and Selection*, 17(1), 107–112.
- Emms DM & S Kelly (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*: 16:157.
- Fan X & Nagle GT. 1996. Molecular cloning of *Aplysia* neuronal cDNAs that encode carboxypeptidases related to mammalian prohormone processing enzymes. *DNA Cell Biol.* 15, 937–945.
- Fan X, Spijker S, Akalala DBG, Nagle GT. 2000 Neuropeptide amidation: cloning of a bifunctional α -amidating enzyme from *Aplysia*. *Brain Res Mol Brain Res*. 82(1-2), 25-34.

- Fassio G., Modica M.V., Mary L., Zaharias P., Fedosov A.E., Gorson J., Kantor Yu.I., Holford M., Puillandre N. 2019. Venom diversity and evolution in the most divergent cone snail genus *Profundiconus*. *Toxins*, 11, 623.
- Fedosov AE & N Puillandre (2012) Phylogeny and taxonomy of the *Kermia*–*Pseudodaphnella* (Mollusca: Gastropoda: Raphitomidae) genus complex: a remarkable radiation via diversification of larval development. *Systematics and Biodiversity*, 10: 447–477.
- Fedosov AE, Stahlschmidt P, Puillandre N, Aznar-Cormano L & P Bouchet (2017) Not all spotted cats are leopards: evidence for a *Hemilienardia ocellata* species complex (Gastropoda: Conoidea: Raphitomidae). *European Journal of Taxonomy*, 268: 1–20.
- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39: 783–791.
- Folmer O, Black M, Hoeh W, Lutz R, & R. Vrijenhoek (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* 3: 294–299.
- Gerdol M, Cervelli M, Oliverio M, Modica MV. 2018. Piercing Fishes: Porin Expansion and Adaptation to Hematophagy in the Vampire Snail *Cumia reticulata*, *Molecular Biology and Evolution*, 35: 2654–2668.
- Gerozissis K. (2003) Brain insulin: regulation, mechanisms of action and functions. *Cell Mol Neurobiol* 23(1): 1–25.
- Gonzales, D. T. T. & Saloma, C. P. (2014) A bioinformatics survey for conotoxin-like sequences in three turrid snail venom duct transcriptomes. *Toxicon*, 92: 66–74.
- Gorson J, Ramrattan G, Verdes A, Wright EM, Kantor YI, Srinivasan RR, Musunuri R, Packer D, Albano G, Qiu WG, Holford M. (2015). Molecular diversity and gene evolution of the venom arsenal of Terebridae predatory marine snails. *Genome Biology and Evolution*, 7: 1761–1778.
- Haas BJ, et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8: 1494–1512.
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41: 95–98.
- Huang Y, Niu B, Gao Y, Fu L, Li W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*.;26(5): 680–682.
- Iacovache I, Bischofberger M, van der Goot FG. 2010. Structure and assembly of pore-forming proteins. *Curr Opin Struct Biol*. 20(2):241–246.

- Imperial JS, Kantor YI, Watkins M, Heralde FM, Stevenson B, Chen P, et al. (2007) Venomous auger snail *Hastula (Impages) hectica* (Linnaeus, 1758): molecular phylogeny, foregut anatomy and comparative toxinology. *Journal of Experimental Zoology*, 308B: 744-756.
- Jimenez EC, Donevan S, Walker C, Zhou LM, Nielsen J, Cruz LJ, et al. (2002) Conantokin-L, a new NMDA receptor antagonist: determinants for anticonvulsant potency. *Epilepsy research*, 57: 73-80.
- Kaas Q, Yu R, Jin A-H, Dutertre S, Craik DJ. 2012. ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Research*; 40 (Database issue): D325-D330.
- Käll L, Krogh A, Sonnhammer EL. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res*: 429-32. Epub 2007 May 5.
- Kantor YI & JD Taylor (2000) Formation of marginal radular teeth in Conoidea (Neogastropoda) and the evolution of the hypodermic envenomation mechanism. *Journal of Zoology*, 252, 251-262
- Kawashima Y, Nagai H, Ishida M, Nagashima Y, Shiomi K. 2003. Primary structure of echotoxin 2, an actinoporin-like hemolytic toxin from the salivary gland of the marine gastropod *Monoplex echo*. *Toxicon* 42(5): 491–497.
- Keane T, Creevey C, Pentony M, Naughton T, McInerney J. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* 6: 29.
- Kohn AB. 2018. *Conus* Envenomation of Humans: In Fact and Fiction. *Toxins*, 11: 10.
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ & DG Higgins (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.
- Li Q, Barghi N, Lu A, Fedosov AE, Bandyopadhyay PK, AO Lluisma et al. (2017) Divergence of the venom exogene repertoire in two sister species of *Turriconus*. *Genome Biology and Evolution*, 9, 2211-2225.
- Lu A, Yang L, Xu S, & Wang C. (2014) Various conotoxin diversifications revealed by a venomomic study of *Conus flavidus*. *Molecular & cellular proteomics*, 13(1), 105–118.
- Macrander J, Panda J, Janies D, Daly M, Reitzel AM. (2018) Venomix: a simple bioinformatic pipeline for identifying and characterizing toxin gene candidates from transcriptomic data. *PeerJ* 6: e5361.

- Miller MA, Pfeiffer W & T Schwartz (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees". *In: Gateway Computing Environments Workshop (GCE)*, New Orleans, pp. 1-8.
- Olivera BM, Gray W, Zeikus R, McIntosh M, Varga J, Rivier J, et al. (1985) Peptide neurotoxins from fish-hunting Cone snails. *Science*, 230: 1338-1343.
- Olivera BM, Seger J, Horvath MP & AE Fedosov (2015) Prey-capture Strategies of Fish-hunting Cone Snails: Behavior, Neurobiology and Evolution. *Brain, Behavior and Evolution*, 86, 58-74.
- Parker MW, Feil SC. 2005. Pore-forming protein toxins: from structure to function. *Prog Biophys Mol Biol*. 88(1): 91–142
- Peng C, Yao G, Gao BM, Fan CX, Bian C, Wang J, Cao Y, Wen B, Zhu Y, Ruan Z et al. (2016). High-throughput identification of novel conotoxins from the Chinese tubular cone snail (*Conus betulinus*) by multi-transcriptome sequencing. *Gigascience*, 5, 17.
- Petersen TN, Brunak S, von Heijne G & H Nielsen (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 8:785–786.
- Phuong MA, Mahardika GN & ME Alfaro (2016) Dietary breadth is positively correlated with venom complexity in cone snails. *BMC Genomics*, 17: 1-15.
- Phuong, MA, GN Mahardika (2018) Targeted sequencing of venom genes from cone snail genomes improves understanding of conotoxin molecular evolution. *MBE* 35:1210-1224.
- Phuong, MA, ME Alfaro, GN Mahardika, RM Marwoto, RE Prabowo, T von Rintelen, PWH Vogt, JR Hendricks, N Puillandre (2019) Lack of signal for the impact of conotoxin gene diversity on speciation rates in cone snails. *Syst Biol* 68:781-796.
- Podobnik M, Anderluh G. 2017. Pore-forming toxins in Cnidaria. *Semin Cell Dev Biol*. 72:133–141.
- Prashanth JR, Brust A, Jin A-H, Alewood PF, Dutertre S & RJ Lewis (2014) Cone snail venomomics: from novel biology to novel therapeutics. *Future Med Chem.*, 6: 1659-1675.
- Puillandre N, Fedosov AE & YI Kantor (2016) Systematics and Evolution of the Conoidea. *In: P. Gopalakrishnakone (Ed), Evolution of Venomous Animals and Their Toxins*. Springer, pp. 1-32.
- Puillandre N, Kantor YI, Sysoev AV, Couloux A, Meyer C, Rawlings T, et al. (2011) The dragon tamed? A molecular phylogeny of the Conoidea (Gastropoda). *Journal of Molluscan Studies*: 77: 259-272.
- Puillandre N, Koua D, Favreau P, Olivera BM & R Stocklin (2012) Molecular phylogeny, classification and evolution of conopeptides. *Journal of Molecular Evolution*, 74: 297-309.

- Puillandre N, Sysoev AV, Olivera BM, Couloux A & P Bouchet (2010) Loss of planktotrophy and speciation: geographical fragmentation in the deep-water gastropod genus *Bathytoma* (Gastropoda, Conoidea) in the western Pacific. *Systematics and Biodiversity*, 8: 371-394.
- Rivera-de-Torre E, Garcia-Linares S, Alegre-Cebollada J, Lacadena J, Gavilanes JG, Martinez-del-Pozo A . 2016. Synergistic action of actinoporin isoforms from the same sea anemone species assembled into functionally active heteropores. *J Biol Chem*. 291(27):14109–14119
- Robinson SD & H Safavi-Hemami (2017) Venom peptides as pharmacological tools and therapeutics for diabetes. *Neuropharmacology*, 127: 79-86.
- Robinson SD, Li Q, Lu A, Bandyopadhyay PK, Yandell M, Olivera BM & Safavi-Hemami H. (2017) The Venom Repertoire of *Conus gloriamaris* (Chemnitz 1777), the Glory of the Sea. *Mar Drugs* 15, 145
- Safavi-Hemami H, Brogan SE & BM Olivera (2019) Pain therapeutics from cone snail venoms: From Ziconotide to novel non-opioid pathways. *Journal of Proteomics*, 190: 12-20.
- Safavi-Hemami H, Gajewiak J, Karanth S, Robinson SD, Ueberheide B, Douglass AD, Schlegel A, Imperial JS, Watkins M, Bandyopadhyay PK, Yandell M, Li Q, Purcell AW, Norton RS, Ellgaard L, Olivera BM. (2015) Specialized insulin is used for chemical warfare by fish-hunting cone snails. *PNAS* 112, 1743-1748.
- Safavi-Hemami H, Lu A, Robinson SD, Li Q, Fedosov AE, Biggs J, et al. (2016) Venom insulins of cone snails diversify rapidly and track prey taxa. *Molecular Biology and Evolution*, 33: 2924-2934.
- Schmieder R & R Edwards (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27: 863-864.
- Shimek RL & AJ Kohn (1981) Functional morphology and evolution of the toxoglossan radula. *Malacologia*, 20: 423-438.
- Smit AB, van Kesteren RE, Li KW, Van Minnen J, Spijker S, Van Heerikhuizen H, Geraerts WP (1998) Towards understanding the role of insulin in the brain: lessons from insulin-related signaling systems in the invertebrate brain. *Progress in Neurobiology* 54(1):35-54
- Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22: 2688-2690.
- Sysoev AV & YI Kantor (1989) Anatomy of molluscs of genus *Splendrilia* (Gastropoda: Toxoglossa: Turridae) with description of two new bathyal species of the genus from New Zealand. *New Zealand Journal of Zoology*, 16: 205-214.

- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, 30(12): 2725–2729.
- Taylor JD, Kantor YI & AV Sysoev, (1993) Foregut anatomy, feeding mechanisms and classification of the Conoidea (= Toxoglossa)(Gastropoda). *Bulletin of the Natural History Museum of London (Zoology)*, 59: 125-170.
- Terlau H & BM Olivera (2004) *Conus* Venoms: A Rich Source of Novel Ion Channel-Targeted Peptides. *Physiological reviews*, 84: 41-68.
- Wagner S & D Wagner 2007. Comparing Clusterings - An Overview.
- Watanabe Y, Tateno H, Nakamura-Tsuruta S, Kominami J, Hirabayashi J, Nakamura O, Watanabe T, Kamiya H, Naganuma T, Ogawa T, Naudé RJ, Muramoto K. (2009) The function of rhamnose-binding lectin in innate immunity by restricted binding to Gb3. *Developmental & Comparative Immunology*, 33(2):187-97.
- Watkins M, Hillyard DR, Olivera BM. 2006. Genes expressed in a turrid venom duct: divergence and similarity to conotoxins. *J Mol Evol*, 62: 247-256.

Captions

Figure 1. Phylogeny and morphology of the genus *Clavus*. **A.** Phylogenetic tree of the Superfamily Conoidea (After Puillandre et al. 2016). **B.** Relationships of the Philippine species of the *Clavus canalicularis* species complex. **C.** Marginal radular teeth of *Conus (Pionoconus) circumcissus* representing a hypodermic type. **D.** Radula of *Clavus canalicularis* with distinct comb-like lateral teeth. **E.** Live *Clavus canalicularis* (Photo – courtesy of David Massemin).

Figure 2. Overview of *Clavus* spp. Venom gland transcriptomes. **A – C.** RBH analysis between conspecific specimens and the two species of *Clavus*. OX – percent identity, OY – RBH pair count. **A.** *Cv. canalicularis* 1 VERSUS *Cv. canalicularis* 2. **B.** *Cv. davidgilmouri* 1 VERSUS *Cv. davidgilmouri* 2. **C.** *Cv. canalicularis* VERSUS *Cv. davidgilmouri*. **D.** Bar graph of the putative toxin gene superfamilies counts in the *Cv. canalicularis* and *Cv. davidgilmouri* (error bars denote standard deviation, spm=specimen). **E.** Scatterplot of the putative toxin gene superfamilies counts in the *Cv. canalicularis* (x) and *Cv. davidgilmouri* (y). **F.** Bar graph of the putative toxin

gene superfamilies total expression levels in the *Cv. canalicularis* and *Cv. davidgilmouri* (error bars denote standard deviation). **G.** Scatterplot of the putative toxin gene superfamilies total expression levels (log-normalized) in the *Cv. canalicularis* (x) and *Cv. davidgilmouri* (y).

Figure 3. Mature peptide region length distribution. **A.** Violin plot of mature peptide region size in *Conus* and *Clavus* spp. **B.** Mature peptide lengths in the putative *Clavus* toxins with BLAST hits. **C.** Scatterplot of putative *Clavus* toxin gene superfamilies – mature peptide region length versus log10 normalized tpm expression level.

Figure 4. Diversity of the predicted toxins in the drillipeptide superfamily 7. **A.** Venn diagram of the drillipeptide diversity in studied specimens of *Clavus*. **B.** Alignment of 64 predicted complete precursors.

Figure 5. *Clavus* insulins. **A.** Comparative sequence alignment of *Clavus* insulins and their expression values (tpm) showing conserved signal peptides and variable A and B chain regions. Amino acid conservations are denoted by an asterisk (*). Full stops (.) and colons (:) represent a low and high degree of similarity, respectively. **B.** Alignment of the B and A chains of two representative *Clavus* insulins (purple) with insulins sequenced from the marine worm *Platynereis dumerilii* and venom insulins sequenced from worm-hunting cone snails (pink). An alignment of the B and A chains of zebrafish insulin with venom insulins from fish-hunting cone snails is shown for comparison. Cysteines are shown in yellow and disulfide frameworks are depicted as connecting lines. Amino acid code: O=hydroxyproline, y=y-carboxylated glutamate, #: C-terminal amidation, **C.** Bayesian phylogenetic trees of the gastropod insulins. Left: based on the analysis of entire precursor; Right: based on the mature peptide region. Diameter of black circles proportional to support (as Posterior probability 0.50 – 1) of corresponding nodes. Thickened lines denote segment of the tree corresponding to conoidean venom insulins. Colored segments as: brown – heterobranchia, dark-blue – *Charonia tritonis* (Tonnoidea), dark green – *Conus* nerve ring, yellow – Turridae, blue – *Clavus* spp., light-green – *Conus* venom gland.

Figure 6. Bayesian phylogenetic tree of neogastropod porins. Coluporins highlighted with orange, Conoporins - with blue, *Clavus* porin clusters - with green. Nodes with the Bayesian PP support > 0.75 marked with grey circles, and with the PP > 0.9 - with black circles.

Supplementary data.

Supplementary Table 1. RNAseq data sets used for this analysis. The number of reads, read length and quality were calculated for samples before and after quality/adaptor trimming

Supplementary Table 2. Summary of assembly and annotation statistics for the 4 RNAseq data sets.

Supplementary Table 3. Statistics of the tested cluster sets.

Supplementary Table 4. Numbers of identical clusters in pairwise comparison of the inferred cluster sets.

Supplementary Table 5. Pearson's correlation coefficients for the intra- and interspecific comparisons of inferred venom components.

Supplementary Table 6. Shannon's diversity indexes of putative *Clavus* venom components in four studied specimens.

Supplementary Figure 1. Number of different Cys-patterns per cluster in different CD-Hit cluster sets.

Supplementary Figure 2. Gene superfamily distribution by mature peptide region length.

Supplementary Figure 3. Bayesian phylogenetic tree of the 64 predicted gene superfamily 4 toxins identified in *Clavus*. Blue segments correspond to venom components solely found in *Cv. canalicularis*, green – in *Cv. davidgilmouri*, grey – orthogroups, identified by OrthoFinder including transcripts from both species. Diameter of black circles proportional to support (as Posterior probability 0.50 – 1) of corresponding nodes.

Supplementary Figure 4. Alignment of the *Clavus* porins with some porins identified in other taxa of predatory gastropods. Three clusters of porins (plus claviporin-Cv6 separately) showed with different color shading. Signal and pro- regions highlighted and underlined.

Supplementary Figure 5. Alignment of the *Clavus* RBLs carbohydrate recognition domains. The putative secondary structure elements adopted from Ogawa et al. 2011.









